

**LAPORAN PRAKTIKUM UJIAN AKHIR SEMESTER
PRAKTIKUM PENGANTAR *DATA MINING***

**ANALISIS KLUSTER UNTUK MENILIK
KEMERATAAN DAN KUALITAS PENDIDIKAN DI INDONESIA**



Nama Mahasiswa (Nomor Induk Mahasiswa):

AISYAH KHAULA IFRA	(21/483079/PA/21057)
ALLISYA MAHARANI ADINDA WIBOWO	(21/478078/PA/20729)
BRYAN FLORENTINO LEO	(21/473767/PA/20429)
FIKRI KAMALUDDIN	(21/476787/PA/20605)
PARAMITA KUMALA DEVI	(21/480819/PA/20910)
TAUFIK HASAN WAU	(21/473864/PA/20436)

**LABORATORIUM KOMPUTASI MATEMATIKA DAN STATISTIKA
DEPARTEMEN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS GADJAH MADA**

2023

ABSTRAK

Pendidikan merupakan suatu aspek yang penting dalam kehidupan manusia sekaligus salah satu indikator kemajuan sebuah negara. Keberagaman lingkungan alam dan masyarakat di Negara Kesatuan Republik Indonesia menimbulkan berbagai tantangan, salah satunya ialah penyelenggaraan pendidikan yang merata dan berkualitas. Beberapa faktor terhadap isu tersebut ialah keterbatasan sosial dan ekonomi, infrastruktur, jumlah dan kualitas tenaga pengajar, serta kurikulum yang tidak dapat mencapai standar nasional yang ditetapkan oleh pemerintah. Dalam menghadapi tantangan tersebut, pemerintah telah melakukan berbagai upaya, seperti dengan mencetuskan program Bantuan Operasional Sekolah (BOS), Kartu Indonesia Pintar (KIP), beasiswa Bidikmisi, dan lain-lain. Lebih lanjut, penempatan guru di berbagai daerah, peningkatan infrastruktur pendidikan, dan pengembangan kurikulum juga telah dilakukan dalam rangka mengurangi kesenjangan pendidikan di Indonesia.

Penelitian ini ditujukan untuk mengetahui daerah-daerah yang memiliki kualitas pendidikan yang baik dan kurang baik di Indonesia dengan metode *clustering*. Indikator pemerataan dan kualitas pendidikan di Indonesia dirangkum dalam enam belas variabel yang diperoleh dari tabel dinamis Badan Pusat Statistik (BPS). Data seluruh variabel diperoleh pada tahun 2022 dan telah terbagi menurut 34 provinsi di Indonesia. Dengan serangkaian praproses, eksplorasi, dan visualisasi data, diperoleh *dataset* berukuran 34 baris dan 16 kolom. *Clustering* dilakukan dengan dua buah algoritma, yaitu *K-means* dan *K-medoids*. Pada kedua algoritma tersebut, penentuan jumlah kluster optimal dicoba pada tiga metode, yaitu kurva Elbow, *silhouette score*, dan *gap statistic*. Jumlah kluster optimal ialah jumlah kluster yang mengakibatkan patahan ekstrem pada kurva Elbow serta memaksimalkan *silhouette score* dan *gap statistic*. Pada akhirnya, diperoleh tiga kluster dari algoritma *K-means* dan lima kluster dari algoritma *K-medoids* dengan kesimpulan bahwa pemerataan pendidikan berkualitas di Indonesia belum baik.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pendidikan merupakan suatu aspek yang penting dalam kehidupan manusia sekaligus salah satu indikator kemajuan sebuah negara. Menurut Pasal 3 Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional, tujuan pendidikan adalah mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermartabat dalam rangka mencerdaskan kehidupan bangsa, bertujuan untuk berkembangnya potensi peserta didik agar menjadi manusia yang beriman dan bertakwa kepada Tuhan Yang Maha Esa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, dan menjadi warga negara yang demokratis serta bertanggung jawab.

Negara Kesatuan Republik Indonesia yang terbentang dari Sabang sampai Merauke, terdiri atas puluhan provinsi, ratusan perkotaan, dan ribuan pedesaan, memiliki lingkungan alam dan masyarakat yang beragam. Keberagaman ini menimbulkan berbagai tantangan, salah satunya ialah penyelenggaraan pendidikan yang merata dan berkualitas. Beberapa faktor terhadap isu tersebut ialah keterbatasan sosial dan ekonomi, infrastruktur, jumlah dan kualitas tenaga pengajar, serta kurikulum yang tidak dapat mencapai standar nasional yang ditetapkan oleh pemerintah. Selain itu, beberapa daerah juga mempunyai tingkat partisipasi pendidikan yang rendah dan angka putus sekolah yang tinggi. Tidak hanya itu, akses pendidikan juga masih terbatas bagi anak-anak berkebutuhan khusus.

Dalam menghadapi tantangan tersebut, pemerintah telah melakukan berbagai upaya, seperti dengan mencetuskan program Bantuan Operasional Sekolah (BOS), Kartu Indonesia Pintar (KIP), beasiswa Bidikmisi, dan lain-lain. Lebih lanjut, penempatan guru di berbagai daerah, peningkatan infrastruktur pendidikan, dan pengembangan kurikulum juga telah dilakukan dalam rangka mengurangi kesenjangan pendidikan di Indonesia.

Penelitian ini ditujukan untuk mengetahui daerah-daerah yang memiliki kualitas pendidikan yang baik dan kurang baik di Indonesia dengan metode

clustering. Melalui penelitian ini, peneliti mengharapkan peningkatan fokus pemerintah maupun segenap rakyat Indonesia dalam mempertahankan dan terus meningkatkan kualitas pendidikan di seluruh wilayah bangsa Indonesia.

1.2 Rumusan Masalah

1. Apakah pendidikan di Indonesia sudah merata?
2. Daerah mana yang sudah memiliki kualitas pendidikan cukup baik?
3. Daerah mana yang kualitas pendidikannya perlu ditingkatkan?
4. Apa saran bagi para pemangku kebijakan untuk meningkatkan kualitas pendidikan di Indonesia?

1.3 Tujuan Penelitian

1. Mengetahui kondisi kerataan pendidikan di Indonesia.
2. Mengetahui kualitas pendidikan di setiap provinsi di Indonesia.
3. Mengetahui daerah mana yang kualitas pendidikannya perlu ditingkatkan.
4. Memberikan rekomendasi kebijakan pendidikan yang sesuai guna meningkatkan kualitas pendidikan di Indonesia.

1.4 Manfaat Penelitian

1. Mengidentifikasi faktor-faktor yang memengaruhi keberhasilan pendidikan di provinsi-provinsi dengan kualitas pendidikan baik.
2. Membantu pembentukan kebijakan-kebijakan yang baik untuk meningkatkan kualitas pendidikan di provinsi-provinsi dengan kualitas pendidikan kurang baik.
3. Membantu pengembangan strategi perbaikan pendidikan yang lebih efektif, meliputi dari segi akses, tingkat literasi, dan lainnya.
4. Membantu penilaian efektivitas dan dampak kebijakan pendidikan yang ada pada saat ini.
5. Memberikan wawasan tentang perbedaan dan dinamika kondisi pendidikan antarregion di Indonesia.

BAB II

PEMBAHASAN

2.1 Metode Analisis

2.1.1 Pengambilan Data

Dalam penelitian ini, kerataan dan kualitas pendidikan di Indonesia diukur melalui enam belas variabel yang diperoleh dari tabel dinamis pada situs web Badan Pusat Statistik (BPS). Masing-masing variabel diperoleh pada tahun 2022 dan telah terbagi menurut 34 provinsi di Indonesia. Keenam belas variabel tersebut adalah sebagai berikut.

1. Angka Partisipasi Murni (APM), yaitu perbandingan antara siswa usia sekolah tertentu pada jenjang pendidikan terhadap penduduk usia yang sesuai dan dinyatakan dalam persentase. APM terbagi menjadi tiga strata (jenjang) yang diwakili oleh tiga variabel terpisah, yaitu APM tingkat SD, SMP, dan SMA.
2. Angka Partisipasi Kasar (APK), yaitu perbandingan antara siswa pada jenjang pendidikan tertentu terhadap penduduk usia sekolah dan dinyatakan dalam persentase. APK terbagi menjadi tiga strata (jenjang) yang diwakili oleh tiga variabel terpisah, yaitu APK tingkat SD, SMP, dan SMA.
3. Angka Partisipasi Sekolah (APS), yaitu perbandingan antara siswa usia sekolah tertentu yang bersekolah pada berbagai jenjang pendidikan terhadap penduduk kelompok usia sekolah yang sesuai dan dinyatakan dalam persentase. APS terbagi menjadi tiga strata (interval usia) yang diwakili oleh tiga variabel terpisah, yaitu APS usia 7 hingga 12 tahun, 13 hingga 15 tahun, dan 16 hingga 18 tahun.
4. Angka Melek Huruf (AMH), yaitu proporsi penduduk usia 15 tahun ke atas yang mempunyai kemampuan membaca dan menulis huruf latin dan huruf lainnya, tanpa harus mengerti apa yang dibaca atau ditulisnya terhadap penduduk usia 15 tahun ke atas.

5. Tingkat penyelesaian pendidikan (TPP), yaitu persentase penduduk pada umur referensi tertentu yang telah menamatkan jenjang pendidikan sesuai kelompok umurnya.
6. Tingkat partisipasi dalam pembelajaran yang terorganisir (TPPO), yaitu angka partisipasi anak usia 6 tahun (satu tahun sebelum usia resmi masuk Sekolah Dasar) dalam program pendidikan yang terorganisir.
7. Rata-rata lama bersekolah penduduk berusia 15 tahun ke atas (RLS).
8. Proporsi remaja dan dewasa berusia 15 hingga 24 tahun dengan keterampilan Teknologi Informasi dan Komputer (TIK).

Setelah diagregasikan, diperoleh *dataset* mula-mula dengan ukuran 35 baris dan enam belas kolom.

2.1.2 Praproses Data (Bagian 1)

Praproses data adalah proses mengubah data mentah menuju bentuk yang lebih mudah dipahami untuk diproses oleh mesin. Praproses data merupakan tahapan yang sangat penting karena kualitas data memengaruhi keberhasilan dan hasil dari suatu analisis. Dengan melakukan praproses data, format data juga diubah, sehingga dapat mengembalikan hasil yang lebih akurat dan optimal.

Dalam penelitian ini, praproses data terbagi menjadi dua tahap. Praproses data bagian pertama dirincikan sebagai berikut.

1. Perbaikan tipe data

Perbaikan tipe data bertujuan untuk memastikan bahwa data yang digunakan dalam analisis memiliki jenis yang sesuai. Contohnya, jika terdapat atribut yang teridentifikasi sebagai *string*, tetapi seharusnya berjenis numerik, tipe data tersebut perlu diubah agar dapat digunakan dalam analisis matematis yang tepat. Data juga perlu mengalami integrasi, karena data bisa dikumpulkan dari berbagai sumber dengan format pencatatan yang berbeda.

2. Penghapusan data duplikat

Data duplikat merujuk pada baris-baris data yang memiliki nilai yang sama untuk semua atributnya. Keberadaan data duplikat dapat memengaruhi hasil analisis karena menganggap data yang sama sebagai entitas yang berbeda. Oleh karena itu, data duplikat perlu dihilangkan guna menghasilkan informasi yang akurat dan analisis yang efisien.

3. Penghapusan baris yang tidak diperlukan

Penghapusan baris yang tidak diperlukan juga ditujukan agar analisis bersifat efisien.

4. Penanganan nilai hilang

Missing value atau nilai hilang merujuk pada nilai yang tidak ada atau tidak tercatat dalam data. Biasanya, hal ini dapat terjadi karena kesalahan input, data yang tidak lengkap, atau kesengajaan untuk tidak diisi. Menangani nilai hilang sangat penting karena data yang tidak lengkap dapat memengaruhi kualitas dan akurasi hasil analisis data. Selain itu, terdapat beberapa metode analisis yang tidak mengizinkan kumpulan data dengan nilai hilang. Salah satu cara menangani nilai hilang adalah dengan mengisi nilai hilang menggunakan fungsi mean dan median untuk data numerik serta fungsi modus untuk data kategorik.

2.1.3 Eksplorasi dan Visualisasi Data

Sebelum melakukan analisis kluster, perlu dilakukan eksplorasi dan visualisasi data guna memahami karakteristik dan pola yang ada dalam data. Terdapat beberapa teknik yang umum digunakan dalam eksplorasi dan visualisasi data sebelum melakukan *clustering* yang terbagi menjadi analisis univariat dan analisis multivariat.

2.1.3.1. Analisis Univariat

Analisis univariat adalah metode-metode statistika untuk menganalisis data yang hanya melibatkan satu variabel. Pada analisis univariat, variabel tersebut dihitung dan dianalisis secara

terpisah untuk menghasilkan gambaran yang lebih jelas. Salah satu alat yang umum digunakan dalam analisis univariat adalah diagram kotak dan titik (*box plot*).

Pada diagram kotak dan titik, terdapat lima ukuran penting yang dari data, yaitu nilai median (Q_2), nilai minimum, nilai maksimum, nilai kuartil bawah (Q_1), dan nilai kuartil atas (Q_3). Apabila letak median tidak berada di tengah kotak, dapat diketahui bahwa data tersebut memiliki sebaran yang tidak simetris.

2.1.3.2. Analisis Multivariat

Analisis multivariat adalah metode-metode statistika untuk menganalisis data yang melibatkan beberapa variabel. Dalam analisis multivariat, terdapat dua alat yang umum digunakan, yaitu matriks korelasi (*correlation heatmap*) dan *pair plot*.

Correlation heatmap adalah representasi grafis dari korelasi antarvariabel dalam suatu *dataset*, di mana nilai-nilai direpresentasikan oleh warna. Korelasi antara dua buah variabel dapat berkisar dari -1 hingga 1 , di mana -1 menunjukkan korelasi negatif sempurna, 0 menunjukkan tidak ada korelasi, dan 1 menunjukkan korelasi positif sempurna. *Correlation heatmap* dapat digunakan untuk mengidentifikasi pasangan-pasangan variabel yang berkorelasi positif atau negatif. Semakin gelap warnanya, semakin kuat korelasinya.

Di sisi lain, *pair plot* adalah visualisasi yang menampilkan diagram pencar (*scatterplot*) untuk setiap pasang variabel dalam suatu *dataset*. *Pair plot* memberikan gambaran visual tentang pola hubungan antarvariabel dan distribusi individu dari setiap variabel. *Pair plot* memungkinkan identifikasi pola hubungan linear, korelasi positif atau negatif, dan distribusi variabel secara cepat dan efisien.

2.1.4 Praproses Data (Bagian 2)

Praproses data bagian kedua meliputi penskalaan data. Penskalaan data dilakukan untuk mengubah rentang nilai dari variabel agar sesuai dengan keinginan peneliti dan memastikan bahwa seluruh variabel memiliki rentang nilai yang sama agar dapat dibandingkan secara adil. Penskalaan data dapat dilakukan dengan beberapa metode, antara lain:

- penskalaan minimum-maksimum (*min-max scaling*), yaitu mengubah rentang nilai variabel menjadi 0 hingga 1;
- normalisasi (*Z-score normalization*), yaitu mengubah nilai variabel sedemikian hingga mean bernilai 0 dan standar deviasi bernilai 1; dan
- *robust scaling*, yaitu mengubah rentang nilai variabel menjadi -1 hingga 1 dengan mempertimbangkan nilai pencilan (*outlier*).

2.1.5 Clustering

Clustering adalah suatu metode pengelompokan data dari unsur-unsur yang lebih kecil berdasarkan adanya kemiripan satu sama lain. Kemiripan yang menjadi dasar pengelompokan tidak bersifat universal, sehingga ukuran-ukuran penyamannya harus dijabarkan terlebih dahulu oleh peneliti. *Clustering* dapat digunakan untuk mengidentifikasi pola atau struktur dalam data yang tidak terlihat. Terdapat beberapa metode *clustering*, seperti *K-means clustering* dan *K-medoids clustering*.

K-means clustering adalah *clustering* berbasis sentroida, di mana jarak antara setiap titik data dan sentroida dihitung untuk menentukan kluster yang akan ditempati oleh setiap titik data, sedangkan *K-medoids clustering* adalah *clustering* berbasis sentroida yang menggunakan medoida sebagai pusat kluster. Medoida adalah titik data dalam kluster yang memiliki jarak rata-rata terkecil dengan semua titik data lain dalam kluster tersebut. *K-means clustering* dan *K-medoids clustering* dapat digunakan dalam berbagai aplikasi, seperti segmentasi pasar dan pengenalan pola.

Kurva Elbow, *silhouette score*, dan *gap statistic* adalah tiga teknik yang dapat digunakan untuk menentukan jumlah kluster yang optimal dan

mengevaluasi kualitas kluster. Dalam kurva Elbow, dipetakan nilai inersia (*sum squared distance*) untuk setiap jumlah kluster dan dipilih jumlah kluster yang menunjukkan penurunan inersia yang signifikan. Dalam *silhouette score*, dihitung koefisien *silhouette* untuk setiap jumlah kluster dan dipilih jumlah kluster yang menghasilkan koefisien *silhouette* tertinggi. Sementara, dalam *gap statistic*, dibandingkan nilai inersia data yang sesungguhnya dengan nilai inersia data yang dihasilkan dari beberapa iterasi *bootstrap*, lalu dipilih jumlah kluster yang menghasilkan *gap* terbesar antara kedua nilai inersia tersebut.

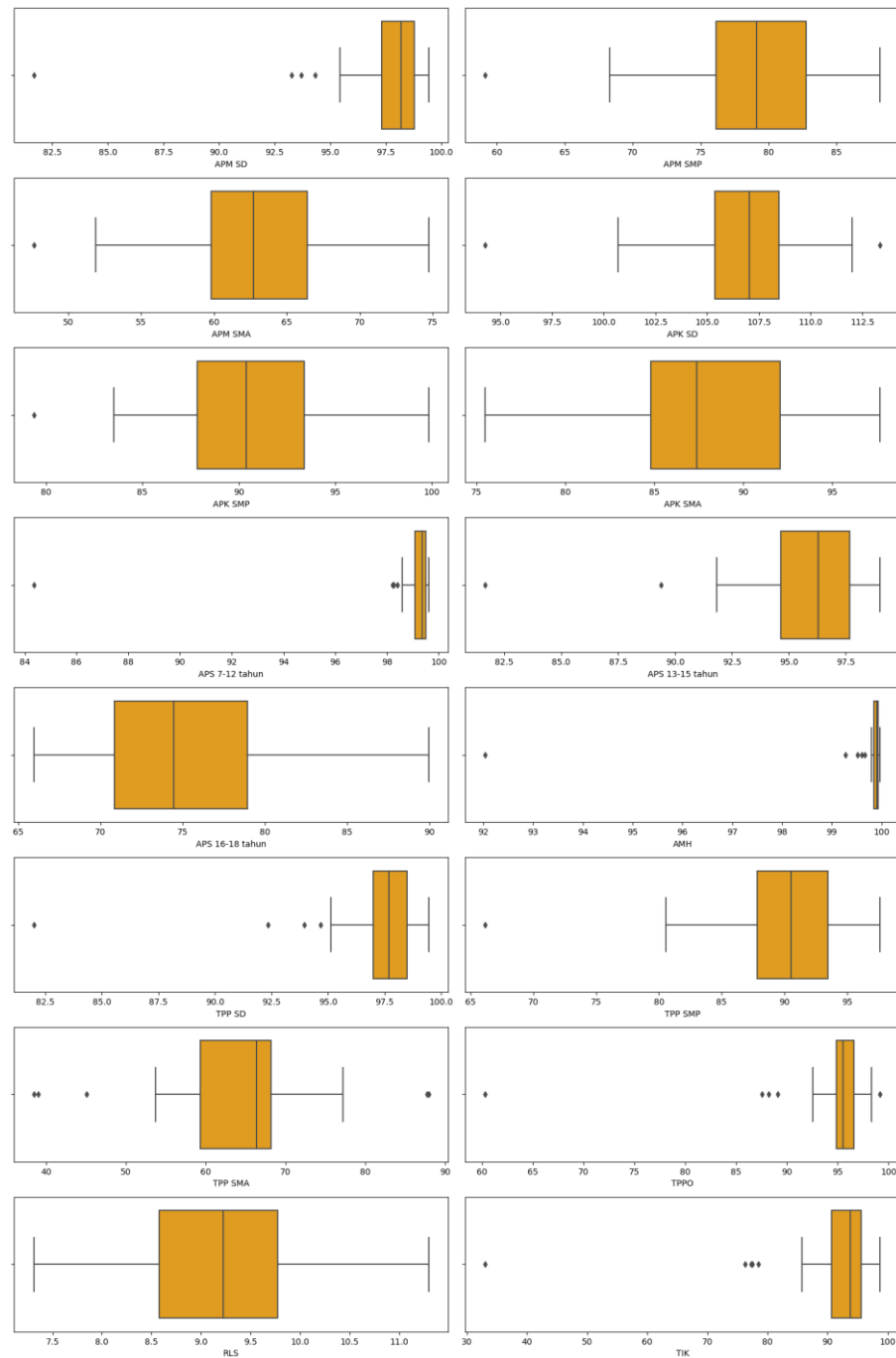
2.2 Analisis

2.2.1 Praproses Data (Bagian 1)

Praproses dilakukan terhadap *dataset* dengan mengubah tipe data, memeriksa dan mengatasi nilai-nilai hilang dan data duplikat, serta membuang variabel yang tidak dibutuhkan. Diperoleh bahwa tipe data mula-mula untuk semua variabel adalah *string*, sehingga dilakukan pengubahan menjadi *float*. Tidak terdapat data hilang dan duplikat. Dibuang baris dengan indeks “INDONESIA” karena tidak mewakili tingkat administratif provinsi.

2.2.2 Eksplorasi dan Visualisasi Data

Eksplorasi dan visualisasi data terbagi menjadi dua bagian, yaitu univariat dan multivariat. Secara univariat, eksplorasi dan visualisasi diawali dengan pembentukan diagram kotak dan titik (*boxplot*) untuk mengamati persebaran data dan keberadaan pencilan dari setiap atribut.

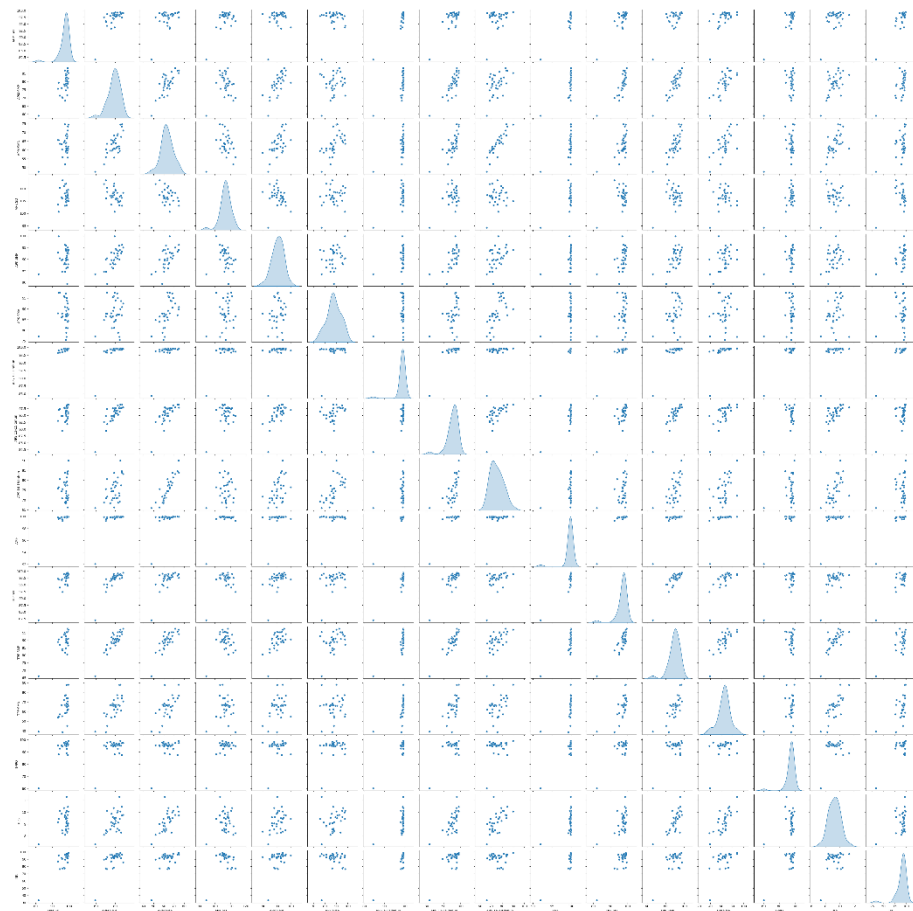


Gambar 1. Diagram kotak dan titik dari setiap atribut.

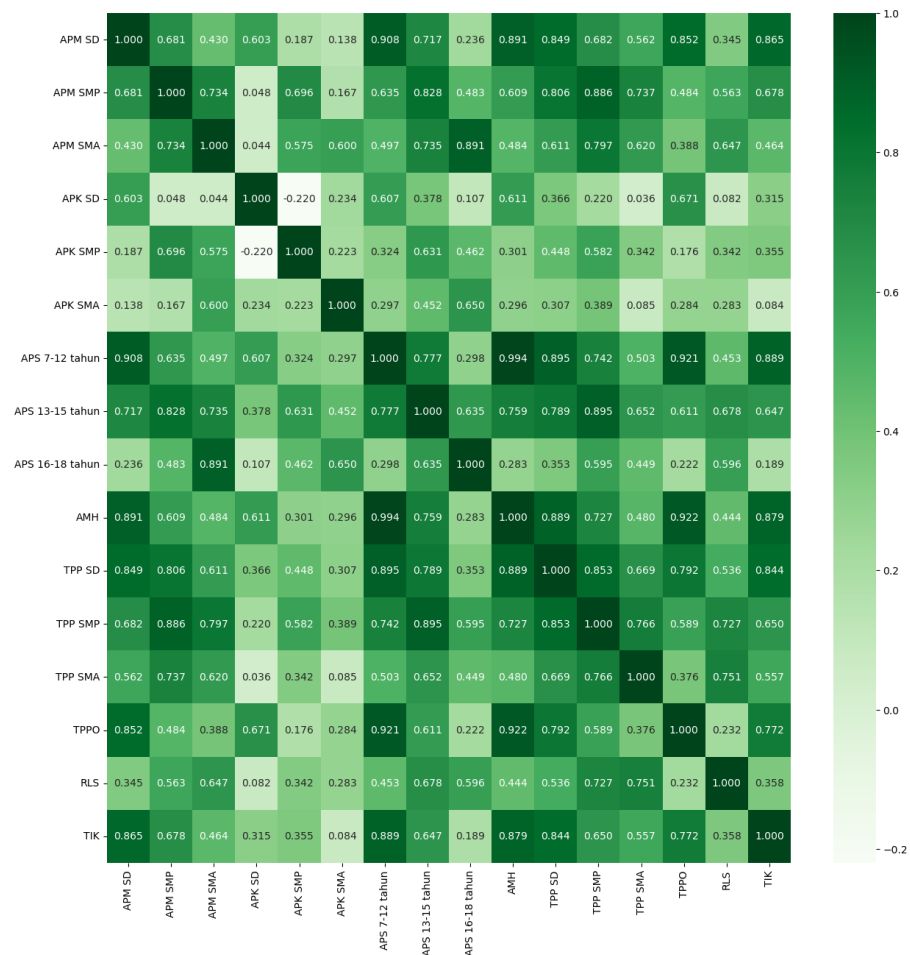
Data variabel APM SD, APS 7-12 tahun, APS 13-15 tahun, AMH, TPP SD, TPP SMP, TPPO, dan TIK mempunyai distribusi menjurai ke bawah atau menceng ke kiri dengan beberapa pencilan. Artinya, pada variabel-variabel tersebut, data dengan nilai tinggi lebih menyebar daripada nilai rendah. Sementara, distribusi data variabel APM SMP dan

APK SD agak menjurai ke bawah. Data variabel APM SMA, APK SMP, APK SMA, TPP SMA, dan RLS mempunyai distribusi yang mendekati simetris karena persebaran data yang hampir merata antara nilai tinggi dan nilai rendah. Variabel APS 16-18 mempunyai distribusi menjurai ke atas atau menceng ke kanan, yang artinya pada variabel tersebut, data dengan nilai rendah lebih menyebar dibandingkan nilai tinggi.

Untuk melihat korelasi antar atribut dalam data, akan digunakan diagram pencar dan *correlation heatmap* atau matriks korelasi.



Gambar 2. Diagram pencar dari setiap atribut.



Gambar 3. Matriks korelasi.

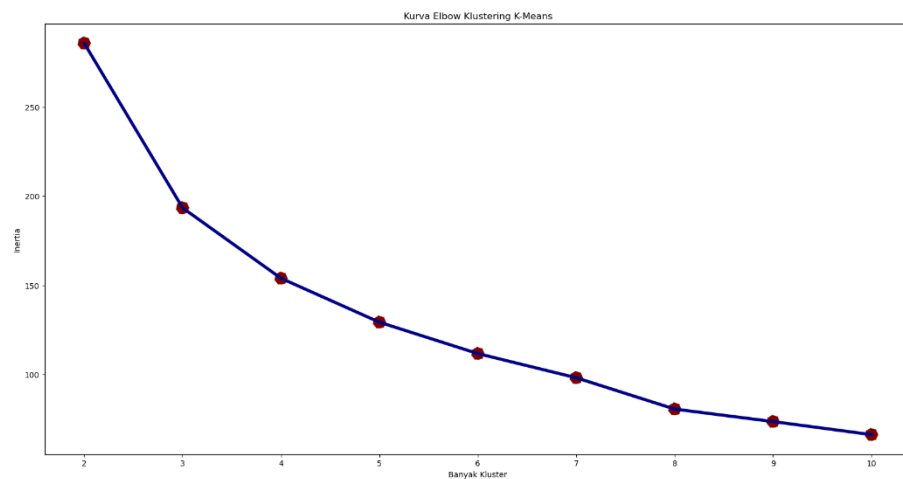
Pada matriks korelasi, semakin gelap warna sel, semakin tinggi korelasi antara dua atribut yang terlibat. Diamati bahwa korelasi positif terbesar terdapat di antara variabel AMH dan APS 7-12 tahun, yaitu senilai 0,994, yang berarti bahwa terdapat hubungan berbanding lurus yang sangat kuat antara Angka Melek Huruf dengan Angka Partisipasi Sekolah usia 7 hingga 12 tahun. Korelasi negatif terbesar terdapat di antara variabel APK SD dan APK SMP, yaitu senilai -0,220, yang berarti bahwa terdapat hubungan berbanding terbalik yang lemah antara Angka Partisipasi Kasar tingkat SD dengan SMP.

2.2.3 Praproses Data (Bagian 2)

Praproses kedua meliputi standarisasi data, sehingga seluruh nilai data berkisar antara 0 dan 1.

2.2.4 Clustering K-Means

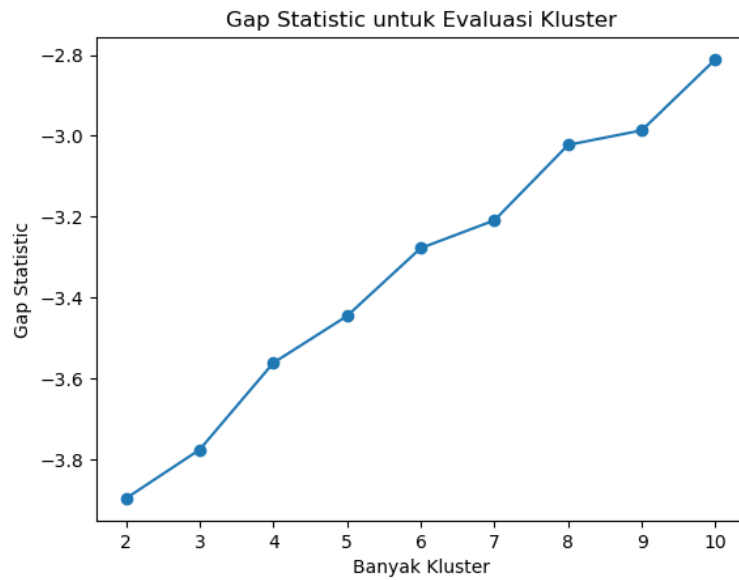
Sebelum dikelompokkan ke dalam kluster, ditentukan jumlah kluster, k , yang tepat untuk *dataset*. Penentuan nilai k menggunakan kurva Elbow, *silhouette score*, dan *gap statistic*. k terbaik ialah nilai yang memberikan patahan pada kurva Elbow, memaksimalkan *silhouette score*, dan memaksimalkan *gap statistic*.



Gambar 4. Kurva Elbow untuk clustering K-means.

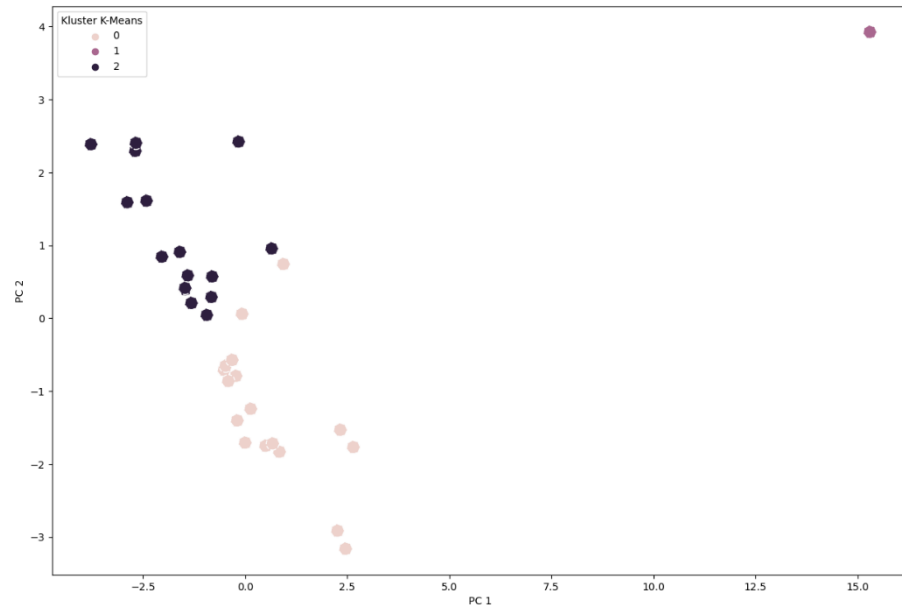
	0
2	0.735188
3	0.248019
4	0.227902
5	0.216309
6	0.214098
7	0.209833
8	0.216633
9	0.190085
10	0.195100

Gambar 5. Silhouette score untuk clustering K-means.



Gambar 6. Gap statistic untuk clustering K-means.

Berdasarkan kurva Elbow, *silhouette score*, dan *gap statistic*, nilai k terbaik secara berturut-turut ialah tiga, dua, dan sepuluh. Ditentukan $k = 3$. Visualisasi ketiga kluster yang terbentuk adalah sebagai berikut.



Gambar 7. Visualisasi ketiga kluster menurut clustering K-means.

Provinsi-provinsi anggota ketiga kluster beserta ciri-ciri numerik menurut ukuran agregasi adalah sebagai berikut.

- Kluster 0 (17 provinsi)
 - Provinsi:

Jambi, Sumatera Selatan, Lampung, Kepulauan Bangka Belitung, Jawa Barat, Jawa Tengah, Banten, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat
 - Ciri numerik:

Nilai rata-rata semua variabel berada pada tingkat menengah, kecuali APK SD, TPPO, dan TIK pada tingkat tinggi.
- Kluster 1 (1 provinsi)
 - Provinsi:

Papua
 - Ciri numerik:

Nilai rata-rata semua variabel berada pada tingkat rendah.
- Kluster 2 (16 provinsi)
 - Provinsi:

Aceh, Sumatera Utara, Sumatera Barat, Riau, Bengkulu, Kepulauan Riau, DKI Jakarta, DI Yogyakarta, Jawa Timur, Bali, Nusa Tenggara Barat, Kalimantan Timur, Kalimantan Utara, Maluku, Maluku Utara, Papua Barat
 - Ciri numerik:

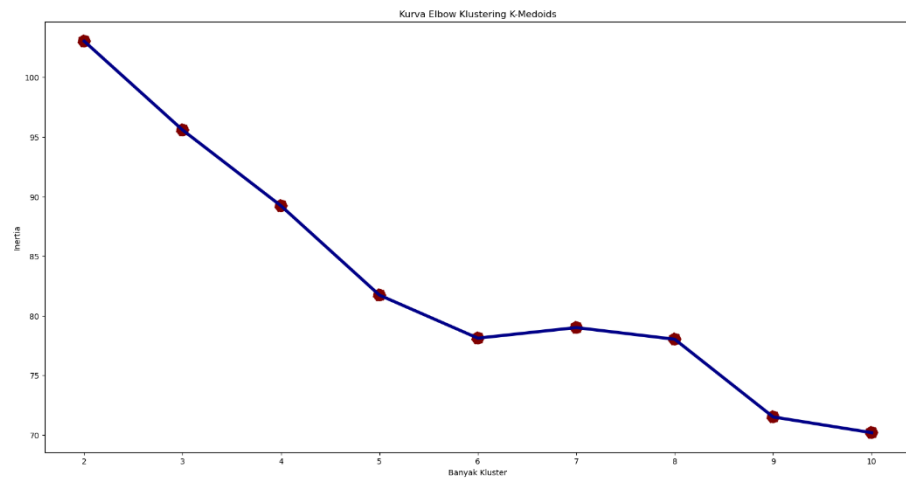
Nilai rata-rata semua variabel berada pada tingkat tinggi, kecuali APK SD, TPPO, dan TIK pada tingkat menengah.

Dengan demikian, menggunakan *clustering K-means*, disimpulkan bahwa mayoritas provinsi di Indonesia telah mempunyai Angka Partisipasi Kasar tingkat Sekolah Dasar (APK SD), tingkat partisipasi dalam pembelajaran yang terorganisasi (TPPO), dan proporsi pemuda dengan keterampilan TIK yang tinggi, sedangkan kualitas yang diukur menggunakan variabel-variabel lainnya berada pada tingkat menengah.

Provinsi Papua adalah provinsi dengan kerataan dan kualitas pendidikan yang paling tidak baik.

2.2.5 Clustering K-Medoids

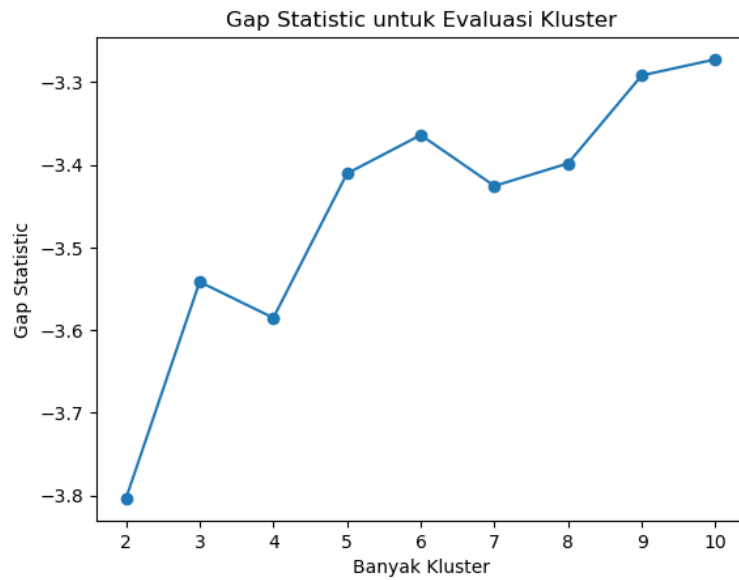
Serupa dengan *clustering K-means*, sebelum dikelompokkan ke dalam kluster, ditentukan jumlah kluster, k , yang tepat untuk *dataset* menggunakan kurva Elbow, *silhouette score*, dan *gap statistic*.



Gambar 8. Kurva Elbow untuk clustering K-medoids.

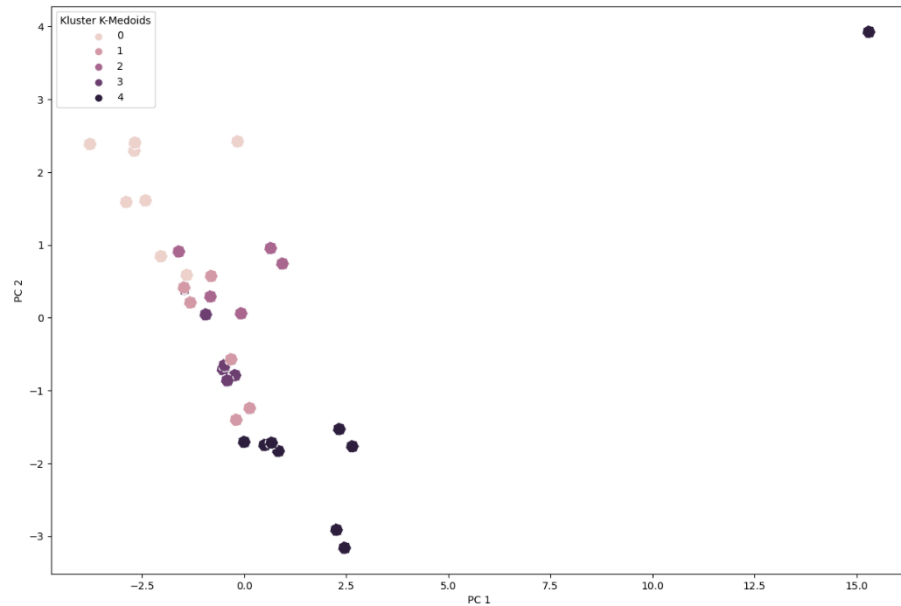
	0
2	0.138903
3	0.088490
4	0.062041
5	0.052432
6	0.054814
7	-0.043661
8	-0.101522
9	-0.063462
10	-0.087223

Gambar 9. Silhouette score untuk clustering K-medoids.



Gambar 10. Gap statistic untuk clustering K-medoids.

Berdasarkan kurva Elbow, *silhouette score*, dan *gap statistic*, nilai k terbaik secara berturut-turut ialah tiga, dua, dan sepuluh. Ditentukan $k = 3$. Visualisasi ketiga kluster yang terbentuk adalah sebagai berikut.



Gambar 11. Visualisasi kelima kluster dari clustering K-medoids.

Provinsi-provinsi anggota kelima kluster beserta ciri-ciri numerik menurut ukuran agregasi adalah sebagai berikut.

- Kluster 0 (8 provinsi)
 - Provinsi:

Aceh, Sumatera Utara, Sumatera Barat, Kepulauan Riau, DI Yogyakarta, Bali, Kalimantan Timur, Kalimantan Utara
 - Ciri numerik:

APM SMP, APM SMA, APK SMP, APK SMA, APS seluruh rentang usia, dan TPP seluruh jenjang tertinggi. APK SD terendah.
- Kluster 1 (6 provinsi)
 - Provinsi:

Jambi, Bengkulu, Sulawesi Selatan, Sulawesi Tenggara, Maluku, Maluku Utara
 - Ciri numerik:

APK SD tertinggi. Nilai variabel-variabel lain cenderung menengah ke tinggi.
- Kluster 2 (5 provinsi)
 - Provinsi:

Riau, DKI Jakarta, Sulawesi Utara, Sulawesi Tengah, Papua Barat
 - Ciri numerik:

APM ketiga jenjang terendah. RLS tertinggi. Nilai variabel-variabel lain cenderung menengah ke rendah.
- Kluster 3 (6 provinsi)
 - Provinsi:

Lampung, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, Nusa Tenggara Timur
 - Ciri numerik:

APM SD, AMH, TPPO, dan TIK tertinggi. Nilai variabel-variabel lain cenderung tinggi atau rendah.
- Kluster 4 (9 provinsi)
 - Provinsi:

Sumatera Selatan, Kepulauan Bangka Belitung, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Gorontalo, Sulawesi Barat, Papua

- Ciri numerik:

APM SMP, APM SMA, APK SMP, APK SMA, APS seluruh rentang usia, AMH, TPP seluruh jenjang, TPPO, RLS, dan TIK terendah.

Dengan demikian, menggunakan *clustering K-medoids*, disimpulkan bahwa mayoritas provinsi di Indonesia telah mempunyai Angka Partisipasi Murni (APM) dan Angka Partisipasi Kasar (APK) jenjang SMP dan SMA, Angka Partisipasi Sekolah (APS) untuk rentang usia 7 hingga 18 tahun, Tingkat Penyelesaian Pendidikan (TPP) untuk jenjang SD hingga SMA, tingkat partisipasi dalam pembelajaran yang terorganisasi (TPPO), rata-rata lama sekolah pemuda, dan proporsi pemuda dengan keterampilan TIK yang rendah. Kerataan dan kualitas pendidikan yang baik sudah terdapat pada enam provinsi dalam kluster 3, yaitu Provinsi Lampung, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, dan Nusa Tenggara Timur.

BAB III

PENUTUP

3.1 Kesimpulan

Berdasarkan *clustering K-means*, kerataan dan kualitas pendidikan provinsi-provinsi di Indonesia terbagi menjadi tiga kluster. Tujuh belas dari 34 provinsi telah mempunyai Angka Partisipasi Kasar tingkat Sekolah Dasar (APK SD), tingkat partisipasi dalam pembelajaran yang terorganisasi (TPPO), dan proporsi pemuda dengan keterampilan TIK yang tinggi, sedangkan kualitas yang diukur menggunakan variabel-variabel lainnya berada pada tingkat menengah. Provinsi Papua adalah provinsi dengan kerataan dan kualitas pendidikan yang paling tidak baik.

Di sisi lain, menurut *clustering K-medoids*, kerataan dan kualitas pendidikan provinsi-provinsi di Indonesia terbagi menjadi lima kluster. Sembilan dari 34 provinsi telah mempunyai Angka Partisipasi Murni (APM) dan Angka Partisipasi Kasar (APK) jenjang SMP dan SMA, Angka Partisipasi Sekolah (APS) untuk rentang usia 7 hingga 18 tahun, Tingkat Penyelesaian Pendidikan (TPP) untuk jenjang SD hingga SMA, tingkat partisipasi dalam pembelajaran yang terorganisasi (TPPO), rata-rata lama sekolah pemuda, dan proporsi pemuda dengan keterampilan TIK yang rendah. Kerataan dan kualitas pendidikan yang baik sudah terdapat pada enam provinsi, yaitu Provinsi Lampung, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, dan Nusa Tenggara Timur.

Dengan demikian, diperoleh bahwa pendidikan berkualitas di Indonesia belum merata.

3.2 Saran

Dari kesimpulan yang telah dipaparkan, beberapa saran yang dapat membantu dalam proses pemerataan dan peningkatan kualitas pendidikan adalah sebagai berikut.

1. Pembangunan sekolah-sekolah bermutu tinggi di provinsi-provinsi dengan kualitas pendidikan rendah, terutama daerah-daerah terpencil.

2. Peningkatan kualitas dan kuantitas guru-guru di Indonesia dan pemberian gaji yang sepadan.
3. Perbaikan akses menuju sekolah serta pemerataan akses listrik dan sinyal agar masyarakat lebih mudah memperoleh informasi.
4. Pemantauan terhadap daerah-daerah terpencil yang belum mendapatkan pendidikan dengan semestinya.

Selain itu, penelitian-penelitian yang akan datang dapat melibatkan variabel-variabel yang lebih serupa dengan indikator-indikator pendidikan yang ditentukan oleh pemerintah serta mencoba metode analisis lain, seperti klasifikasi atau *clustering* dengan algoritma lain.

DAFTAR PUSTAKA

- Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2018). *INDIKATOR PENDIDIKAN DI INDONESIA TAHUN 2017/2018*. Jakarta: Pusat Data dan Statistik Pendidikan dan Kebudayaan.
- Muttaqien, F. (2022, September 28). *4 Langkah Data Preprocessing Agar Data Lebih Mudah Dibaca*. Diambil kembali dari EKRUT media: <https://www.ekrut.com/media/data-preprocessing>
- Sikken, A. S. (2021, Juli 25). *Mengenal Box Plot dan Fungsinya Sebagai Media Visualisasi Data*. Diambil kembali dari Laboratorium Analisis Data dan Rekayasa Kualitas Universitas Brawijaya: https://lab_adrk.ub.ac.id/id/mengenal-box-plot-dan-fungsinya-sebagai-media-visualisasi-data/
- Tampubolon, H. D., Suhada, Safii, M., Solikhun, & Suhendro, D. (2021). Penerapan Algoritma K-Means dan K-Medoids Clustering untuk Mengelompokkan Tindak Kriminalitas Berdasarkan Provinsi. *JURNAL ILMU KOMPUTER DAN TEKNOLOGI*, 6-12.

LAMPIRAN

Tautan video rekaman presentasi:

<https://youtu.be/RTBgbJtX1x0>